2004

# Flexibility and sequence variability in proteins

Haihong Liao
*San Jose State University*

# FLEXIBILITY AND SEQUENCE VARIABILITY IN PROTEINS

A Thesis

Presented to

The Faculty of the Department of Chemistry

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Haihong Liao

December 2004

APPROVED FOR THE DEPARTMENT OF CHEMISTRY

Dr. Brooke Lustig

Dr. Elaine D. Collins

Dr. Steve Branz

Dr. Pamela Stacks

APPROVED FOR THE UNIVERSITY

# ABSTRACT

## Flexibility and Sequence Variability in Proteins

by Haihong Liao

Approaches to determine protein structural features from protein sequence include molecular mechanics and motif recognition. Here one correlates a Shannon information entropy defined as sequence entropy with respect to the flexibility of native globular proteins as described by inverse packing density. For individual proteins and the accompanying aggregate set a strong linear correlation is observed between the calculated sequence entropy and the corresponding flexibility determined at an associated residue position. Three different hydrophobicity scales were applied to the set of query proteins, and all three sets of query hydrophobicity values share similarity with the corresponding sequence entropy values. There appears strong correlation among sequence variability, relative hydrophobicity, and structural flexibility.

## ACKNOWLEDGEMENTS

Dr. Brooke Lustig (Research Director)

Dr. Elaine Collins

Dr. Pamela Stacks

Dr. Stephen Branz

Yaojun Luo

David Chiang

William Yeh

My family

TABLE OF CONTENTS

# LIST OF TABLES

LIST OF FIGURES

# INTRODUCTION

## Protein overview

Proteins are linear unbranched polymers of amino acids (Brown, 2002). There are 20 naturally occurring amino acids that serve as building blocks for proteins, each consists of a central carbon atom (the alpha carbon, $C^{\alpha}$) to which is attached a hydrogen atom, an amino group ($-NH_2$), a carboxyl group (-COOH), and a side chain called the R group. The R group varies from one amino acid to another and gives each amino acid specific properties.

A protein can have four distinct levels of structure. The primary structure of the protein is the polypeptide chain of amino acids. The secondary structure refers to the different conformations of the polypeptide focusing exclusively on backbone interactions. Two main types of secondary structure are the $\alpha$-helix and $\beta$-sheet, both of which are stabilized by backbone hydrogen bonds and often involve local interaction. The tertiary structure includes folding the secondary structural components of the polypeptide into a three-dimensional configuration. The tertiary structure is stabilized by various interactions, notably hydrogen bonding between individual amino acids and

hydrophobic forces. This of course has an impact on specific protein function given the strong relationship between structure and function. The quaternary structure involves the association of two or more polypeptide chains.

Proteins are functionally diverse because the component amino acids are chemically diverse, deriving from the varying R groups. Different sequences of amino acids therefore result in different combinations of chemical reactivities. The amino acid compositions of a protein uniquely determine the three-dimensional structure of the protein (e.g., two proteins with the same amino acids sequence will have the same three-dimensional structure). Proteins are known to have many important functions in the cell, such as enzymatic activity, structure, movement, storage, regulation of cellular processes, transport of material, signal transduction, and immune response.

Some 27,112 proteins (since 7-Sep-2004) of known structure are stored in the Protein Data Bank (Bernstein et al., 1977). The common experimental methods for finding protein three-dimensional structures are X-ray diffraction, neutron-diffraction and nuclear magnetic resonance (NMR). However, these methods are slow and costly (often several months or years of lab work), and much slower than DNA and

protein sequencing. Furthermore, for some protein classes such as transmembrane proteins, the experimental methods to determine three-dimensional structures are problematic because the proteins are either difficult to crystallize or too large for NMR. The latter is apparently a result of the difficulty of acquiring and analyzing the large amount of signal associated with such large proteins. This creates interest in algorithms for protein structure prediction. Theoretical understanding of how proteins fold will allow scientists to quickly and reliably predict protein structures, to design proteins not found in nature, and to model the biological function of individual molecules as well as complex pathways (Brown, 2002).

**Sequence homology**

Determining sequence homology provides one of the most powerful tools available for protein structure prediction. Sequence homology can provide information on the function of an entire protein or the segments within it. The basis of such analysis is that functionally and structurally related proteins often have similar sequences. Very often, the tertiary structures of proteins are well conserved during protein evolution (Miyazawa et al., 1993). This is

because the function of a protein may be conserved during protein evolution, and the function of a protein is closely related to its three-dimensional structure.

A particular tertiary structure is important for protein function. The stability of protein tertiary structures can be significantly affected by amino acid substitutions in the primary sequences. The effects of amino acid substitutions depend on the type of replacement. On average, the stability of tertiary structures is less affected by substitutions among amino acids with similar physico-chemical properties than by others (Dayhoff et al., 1978). Therefore, any amino acid mutation that makes protein native structures unstable is generally harmful for a host organism, and is therefore eliminated from a population in the process of evolution. As a result, sequences with related functions should be informative about the ranges of viable substitutions.

Sequence homology makes use of these similarities, comparing the query protein sequence with other sequences in the databases. Several software programs exist for this type of analysis, and the most popular one is BLAST (Basic Local Alignment Search Tool) Altschul et al. (1997). The NCBI (The National Center for Biotechnology Information)

BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. BLAST uses a strategy based on matching sequence fragments by employing a powerful statistical model to find the best local alignments (Althchul et al., 1990). For protein analysis, BLASTP is a program for comparing a protein query sequence with protein databases.

**Structure-based informatics**

So far, the exact relationship between protein sequence and structure is only partially understood (Jones, 2000; Baker & Sali, 2001). One of most successful computational methodologies for predicting protein structure is using sequence alignment in conjunction with molecular modeling (Marti-Renom et al., 2000). There are some protein comparison methods that implement structure identification. One can identify particular folds using multiple sequence alignments combined with structural information predicted from the sequence of the target (Fischer & Eisenberg, 1997). It was also found that multiple alignments of regions of secondary structure are useful in the identification of key hydrophobic residues

when utilizing hydrophobic cluster analysis (Poupon & Mornon, 1999; Gross et al., 2000).

**Hydrophobicity**

Hydrophobicity is a measure of how miscible an amino acid is with respect to water solvent. It plays a key role in protein structure and hydrophobic residues are often conserved during evolution (Miyazawa & Jernigan, 1993). Hydrophobic effects arise because the hydrogen-bonded structure of water forces hydrophobic groups to become buried inside the proteins. The most obvious result of this is that the most hydrophobic residues should be buried in a protein core. Therefore, interiors of proteins tend to contain fewer charged and polar residues, and the more nonpolar residues are found on the surface in contact with water (Engleman et al., 1986). Hydrophobic effects are not true bonds but they are the main determinants of protein three-dimensional structure, and the burial of hydrophobic groups is a significant determinant of stabilization energy for proteins (Kauzmann, 1959).

A number of scales of hydrophobicity have been developed for amino acids. Significant differences exist among the scales (Engleman et al., 1986). The Hopp-Woods

(1981) scale utilized predictions of potential antigenic sites in globular proteins, assuming they are likely to be rich in charged and polar residues. The scale is essentially a hydrophobicity index resulting in nonpolar residues typically being assigned negative values. Hopp-Woods optimized the original Levitt (1976) scales determined from the measured free energy of transfer of individual amino acid from water to ethanol. When lacking experimental information, the Levitt scales were estimated from the proposed relationship between accessible surface area and hydrophobicity. Subsequently Hopp-Woods parameters are optimized with respect to a number of experimentally characterized antigenic determinants.

Engleman, Steitz and Goldman (1986) developed a hydrophobicity scale from the hydrophobic and hydrophilic components of transfer of amino acid side chains from water to a nonaqueous environment. The scales have been specifically developed for amino acids in $\alpha$-helical structures. The free energy of transfer of both the hydrophobic and hydrophilic components of each amino acid from water into a nonaqueous environment, enclosed by a lipid bilayer with a dielectric constant of 2, were assigned. The hydrophobic/hydrophilic component of the free

energy of water-bilayer transfer can be calculated from the surface area of an amino acid side chain in a $\alpha$-helix. Hydrophobic interactions tend to reduce the nonpolar surface area in contact with water. Their approximate magnitude has been obtained by measuring the partitioning of compounds between water and nonpolar solvents. The hydrophobic/hydrophilic free energy thus measured has been shown to correlate linearly with total surface area in contact with water. For aspartic and glutamic acid there are considerations for the energy required to convert the charged side chains to neutral species at pH7 (Engleman & Steitz, 1981). For carboxyl groups, the energy cost must be considered in two stages. There is the energy cost of removing the protonated group from contact with the aqueous environment, approximately 4.3 kcal/mole, and the energy required to protonate the carboxyl group, given by

$$\Delta G = -1.36(pK - 7)$$

Sharp, Honig and coworkers (1991) refined existing experimentally determined residue hydrophobicity values (Fauchere & Pliska, 1983). These experiments calculate changes in free energy for transferring an individual amino acid from octanol to water. Sharp et al. corrected for the

size of side chains. All three hydrophobicity scales (Hopp-Woods, Engleman-Steitz, Sharp-Honig scales) are different in origin and represent a reasonable sample of the many such scales.

**Shannon entropy**

Entropy has important physical implications as the amount of "disorder" of a system. Entropy measures the degree of disorder in the system. In proteins, the entropy can be defined as (Kono & Saven, 2001)

$$S = -\sum_{\substack{protein \\ states}} W(\text{protein state}) \ln W(\text{protein state})$$

Here, "protein states" refers to all the conformational degrees of freedom necessary to specify the state of the folded protein. The probability of a protein being in a particular state is W.

Shannon's entropy is the central feature of information theory sometimes referred to as a measure of uncertainty, thus a measure of the amount of information. Sequence entropy is the relevant Shannon entropy expression for proteins and nucleic acids. It is defined as

$$S_k = -\sum_{j=1,20} P_{jk} \ln P_{jk}$$

where $P_{jk}$ is the probability of observing a particular amino acid j at sequence position k (Kono & Saven, 2001), and the PlnP term is defined as 0 if P=0. The more partitioning of a set of events (i.e. $P_{jk}$ distributes more evenly), the larger is the Shannon entropy (the range of sequence entropy values are from 0 to 3).

One can identify unique protein secondary structures from patterns of variability in amino acid sequence by using information theory (Pilpel & Lancet, 1999). By exploring large-scale sequence space, people have found that sequence entropy values cluster corresponding to a particular fold (Larson et al., 2002). Initially an expression of Shannon entropy to nucleic acid sequence variability was proven useful in identifying DNA control regions (Schneider et al., 1986; Papp et al., 1993). The method was further extended so as to measure amino acid conservation in proteins (Valdar, 2002).

It has been shown that Shannon derived entropies for protein sequence correlate with entropies calculated from local physical parameters, including backbone geometry (Koehl & Levitt, 2002). Conventional generalized chain statistics appear to significantly over represent the magnitude of the entropic penalty associated with loop

closure in RNA and proteins (Lustig et al., 1998; Scalley-Kim et al., 2003). To understand protein stability and function, it is important to understand the interplay between entropy, structure and sequence.

In this research, a large set of aligned protein sequences is generated from a diversified collection of 130 query sequences. For each residue position, sequence entropy is calculated, then compared to the corresponding local flexibility calculated from the three-dimensional structural data of the query sequence. The residue hydrophobicity is also determined for each query protein with three different scales: Sharp-Honig, Hopp-Woods, and Engleman-Steitz. The average hydrophobicity is calculated by summing the residue hydrophobicities for all residues found within an interval of packing density. Then it is compared to the corresponding calculated flexibility for relationships among sequence variability, relative hydrophobicity, and structural flexibility. Clearly, this is an attempt to quantify the previous qualitative conventional wisdom concerning the degree of residue burial being somehow a measure of the residue's conserved nature and hydrophobicity. Therefore it is expected that a small sequence entropy (i.e. limited sequence variability) should

be calculated for a buried and thus relatively inflexible residue sequence position. Such residue positions are expected to correspond to strongly hydrophobic characteristics.

## METHODS

### Packing density

A diverse, well-characterized set of 130 query protein sequences (see Table 1) are compiled from the Protein Data Bank (PDB, 2002) that satisfy the following conditions.

- Protein structures are required to be determined by X-ray analysis. All protein structures determined by NMR and any other methods are excluded.
- Proteins are required to be 85 or more residues.

For each residue of target query proteins, $C^{\alpha}$ packing density is calculated from its associated atomic coordinates (x, y, z) which are available in PDB Database. An optimal radius of $C^{\alpha}$ packing is determined for 9 Å around a given $C^{\alpha}$ residue position. The average volume of an amino acid is about 111 $Å^3$ (Creighton, 1994), and the average radius of an amino acid is about 2.98 Å. Radii of 7, 8, 9, 10, and 11 Å have been utilized to calculate local $C^{\alpha}$ density, and the density data is best distributed with 9 Å as the radius of the density counting sphere. The assumption here is that the inverse of local packing density is correlated with local flexibility (Bahar et al., 1997), the larger the inverse of local packing density, the greater the local flexibility.

For each query protein, the distance D(i,j) between any two residues i and j of the protein is calculated by the following formula.

$$D(i,j) = \sqrt{(x(i) - x(j))^2 + ((y(i) - y(j))^2 + (z(i) - z(j))^2}$$

For each residue, count the number of the residues (including itself) within the 9 Å range of this residue, and then this value is used as the local packing density of the residue. This is done using a Perl program to input the PDB data file (Appendix A).

Perl is a popular computer language that has powerful character processing functions. It is extensively used in bioinformatics. Using Perl programs, text files with large amounts of data such as PDB files can be easily parsed and processed. The information in those files is extracted and manipulated by Perl programs then further calculation and analysis can be done on the data. Two Perl programs are used to expedite the data processing of protein data obtained from PDB and BLASTP (see Appendix A and B). The program pdb2den.pl uses the PDB file as input, then extracts the protein sequence and calculates the density of each residue of the protein. An output file is generated with the protein sequence and all the densities of each residue. The residue

sequence obtained here is used in the BLAST sequence alignment as the query.

**Sequence alignment**

Aligned sequences are generated with BLASTP for each of these protein sequences obtained from a PDB file (Altschul et al., 1997) searching GenBank and made available by National Center of Biotechnology Information (NCBI). Each BLASTP search output is edited and analyzed using the Perl program (Appendix B). If the protein alignment set satisfies the following conditions, the sequence entropy is calculated for each residue by the same Perl program.

- On the BLASTP query page, alignments are set to 100, which will give maximum number of alignments each query returns 100. Almost all significant alignments can use are within first 100 alignments for all query proteins.
- Alignment bit scores are equal to or higher than 40 percent of the highest score, and are not less than 100. The most significant alignments have scores larger than 40 percent of the highest score.

- The number of alignments is not to be less than 10 for a particular query. If the number of alignments is less than 10 for a query, the corresponding alignment set will be discarded because of the likely absence of statistical significance.

- Gapped regions are ignored. A missed residue for a target residue position is not counted in the entropy calculation.

A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, a penalty (gap score) for a gap is deducted from the alignment score. Extension of the gap to encompass additional nucleotides or amino acids is also penalized in the scoring of an alignment.

The raw score, S, for an alignment is calculated by summing the scores for each aligned position and subtracting the scores for gaps (Altschul et al., 1996). In amino acid alignments, the score for an identity or a substitution is given by the specified substitution matrix.

$$S=\sum(\text{identities, mismatches})-\sum(\text{gap penalties})$$

$$\text{Score}=\text{Max}(S)$$

Bit score S′ is derived from the raw alignment score, S, in which the statistical properties of the scoring system used have been taken into account (Altschul et al., 1996; Karlin & Altschul, 1990). A formula is used to convert a raw score S into a normalized score S' expressed in bits:

S'=(lambda*S-lnK)/(ln 2)

where lambda and K are parameters dependent upon the scoring system (substitution matrix and gap costs) employed.

**Sequence entropy**

For a set of peptide or protein sequences, an expression for sequence entropy, $S_k$, at an amino acid position, k, can be expressed as

$$S_k = -\sum_{j=1,20} P_{jk} \ln P_{jk}$$

where the probability, $P_{jk}$, at some amino acid sequence position k is derived from the frequency $f_{jk}$ for an amino acid type j (e.g., Lys) at sequence position k for N aligned residues. To calculate the probability $P_{jk}$, at residue position k for amino acid type j of a protein alignment set, the corresponding number of residues $f_{jk}$ is

counted and divided by the total number of aligned residues, N. Gapped residues are ignored. The range of sequence entropy is from 0 to 3. The calculations for sequence entropy are done using Perl program with the BLASP query result file as input (Appendix B).

**Correlation plots**

For each query protein, a correlation plot of sequence entropy versus $C^{\alpha}$ packing density is obtained by SigmaPlot 5.0. A linear regression is applied to the correlation plot. Using the number of query protein residues and r squared value, the probability P that the observed data could have come from an uncorrelated parent population (Bevington, 1969) is calculated for each protein (Appendix C).

$$P = \frac{1}{\sqrt{\pi}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \int_{|r|}^{1} (1-x^2)^{1/2(\nu-2)} dx$$
$$\nu = N - 2$$

Note r is linear-correlation coefficient, N the number of observations, and $\nu$ refers degrees of freedom. Typically, the probability P is less than 0.001, suggesting the correlation is statistically significant.

The aggregate (i.e., for all 130 alignment sets) sequence entropy versus inverse $C^{\alpha}$ packing density correlation plots are applied to show the relationship between entropy and packing density. Single averaging is done by summing individual residue entropies for a particular $C^{\alpha}$ packing density interval from all 130 protein sets of alignments. Double averaging entails summing over all proteins the averaged entropy per density interval for individual proteins. The two types of averaged sequence entropy are effectively identical.

**Hydrophobicity**

Residue hydrophobicity is obtained for each query protein with three different hydrophobicity scales: Engleman-Steitz, Sharp-Honig, and Hopp-Woods. For each scale, an averaged hydrophobicity is calculated from the set of all residues within a particular density interval. Overlays are applied to three sets of normalized averaged hydrophobicities and single averaged sequence entropy as a function of inverse packing density.

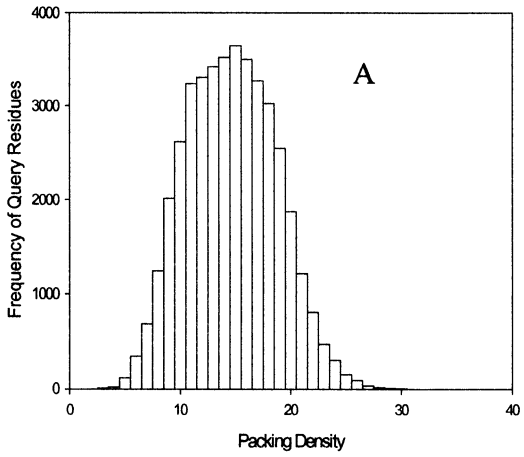
## RESULTS

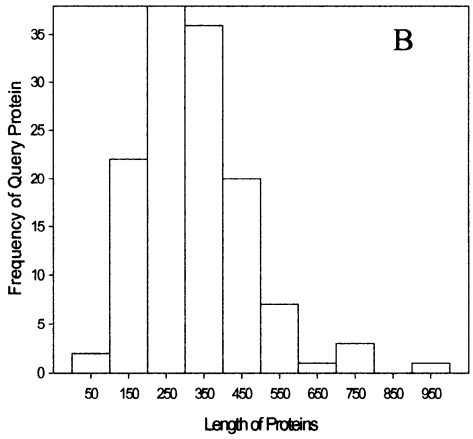
The query proteins' frequency distribution of residues is shown in Figure 1A. The total of 41,632 residues of all 130 proteins with respect to each packing density is slightly right-skewed in the distribution plot. The average, median, mode frequency values per density interval of $C^{\alpha}$ per 9 Å radius, and the range of such values are 14.62, 15, 15 and 0 to 35. The packing densities of 1, 2, 33, and 34 $C^{\alpha}$ atoms within 9 Å radius have minimal occupancy, while the packing density of 15 $C^{\alpha}$ atoms has the largest component of 3661 residues.

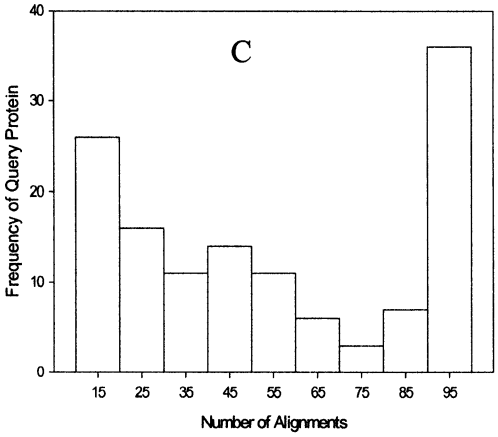
The distribution for the length of query proteins is shown in Figure 1B. The average, median, and mode number of amino acid length per query protein for all 130 proteins are 320.28, 306, and 336. The range of such values is 85 to 901 with proteins 1a1i and 1a32 containing 85 residues and 1bg3 901 residues.

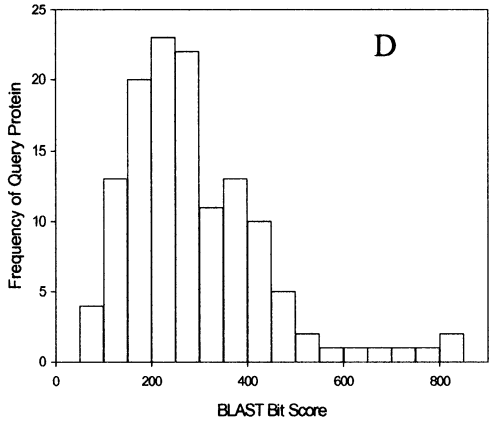
The query sequence alignment by NCBI BLASTP results in a representative distribution of 7143 aligned protein sequences for all 130 proteins. The frequency distribution of the number of alignments is shown in Figure 1C. The average, median, and mode number of alignments per query

**Figure 1**. Frequency distributions for a set of descriptors of query proteins and BLASTP alignments. A) Frequency of total of 41,632 query residues for all 130 proteins with respect to each packing density. Note very small populations are not shown here at packing densities of 31-35. B) Frequency of query proteins with respect to the length of proteins. C) Frequency of query proteins with respect to the number of corresponding alignments determined by BLASTP. There are a total of 7,143 aligned protein sequences for all 130 proteins. D) Frequency of query protein with respect to the lowest BLAST bit score for each corresponding set of alignments.


A
Frequency of Query Residues
4000
3000
2000
1000
0
0
10
20
30
40
Packing Density


B
Frequency of Query Protein
35
30
25
20
15
10
5
0
50
150
250
350
450
550
650
750
850
950
Length of Proteins


C
Frequency of Query Protein
40
30
20
10
0
15
25
35
45
55
65
75
85
95
Number of Alignments


D
Frequency of Query Protein
25
20
15
10
5
0
0
200
400
600
800
BLAST Bit Score

for all 130 proteins are 54.95, 47, and 100. Except for protein 1aq0, which has 101 alignments, the general range of number of alignments is 10 to 100. There are 32 query proteins that have alignments of 100.

The most important criteria in protein BLAST alignment is the bit score. The alignment bit score is calculated from the raw alignment score. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches. The lowest alignment bit scores should be equal to or higher than 40 percent of the highest score and not less than 100. Lower thresholds were explored but proved unreliable in their resulting sequence entropy.

The average, median, mode and the overall range of highest BLASTP bit scores for all 7,143 alignments are 635.1, 594, 662 and 127 to 1793. Protein 1a32 has the highest bit score of 127 and 1bg3 of 1793. The frequency distribution of the 130 protein set of lowest BLASTP bit scores (shown in Figure 1D) is consistent with the right-skewed distribution for a randomized set of BLAST scores (Altschul, 1994). Here the average, median, mode and the overall range of lowest BLASTP bit scores for all 7143

alignments are 293.2, 257,100 and 100 to 831. Protein 5cpv has the lowest bit score of 100 and 1bf2 the highest of 831.

The sequence entropy for each query protein is compared against the inverse of the $C^{\alpha}$ packing density (see Table 1 for summary). For most of proteins (101 proteins out of total 130 proteins), the probability, P, that the observed data could have come from an uncorrelated parent population (Bevington, 1969) is less than 0.001. It is unlikely that such a significant degree of correlation among the set of plots is random in nature.

Examples of protein correlation plots showing sequence entropy for each query residues versus inverse of packing density are shown in Figures 2A through 2C. Representive plots with respect to the corresponding P values are shown for pepsinogen (3psg, 365 residues), dihydrofolate reductase (4dfr, 158 residues.), and oncomodulin (1omd, 107 residues). The respective straight-line fits for all points are $y=13.020x-0.088$ ($P<0.001$), $y=6.064x+0.342$ ($P<0.001$), and $y=4.328x+0.427$ ($P<0.15$), where the respective correlation coefficients are 0.447, 0.274 and 0.141. For all 130 effectively unchanged upon averaging. For pepsinogen, dihydrofolate reductase, and oncomodulin, the respective

Table 1. Summary of linear correlation results for 130 proteins.

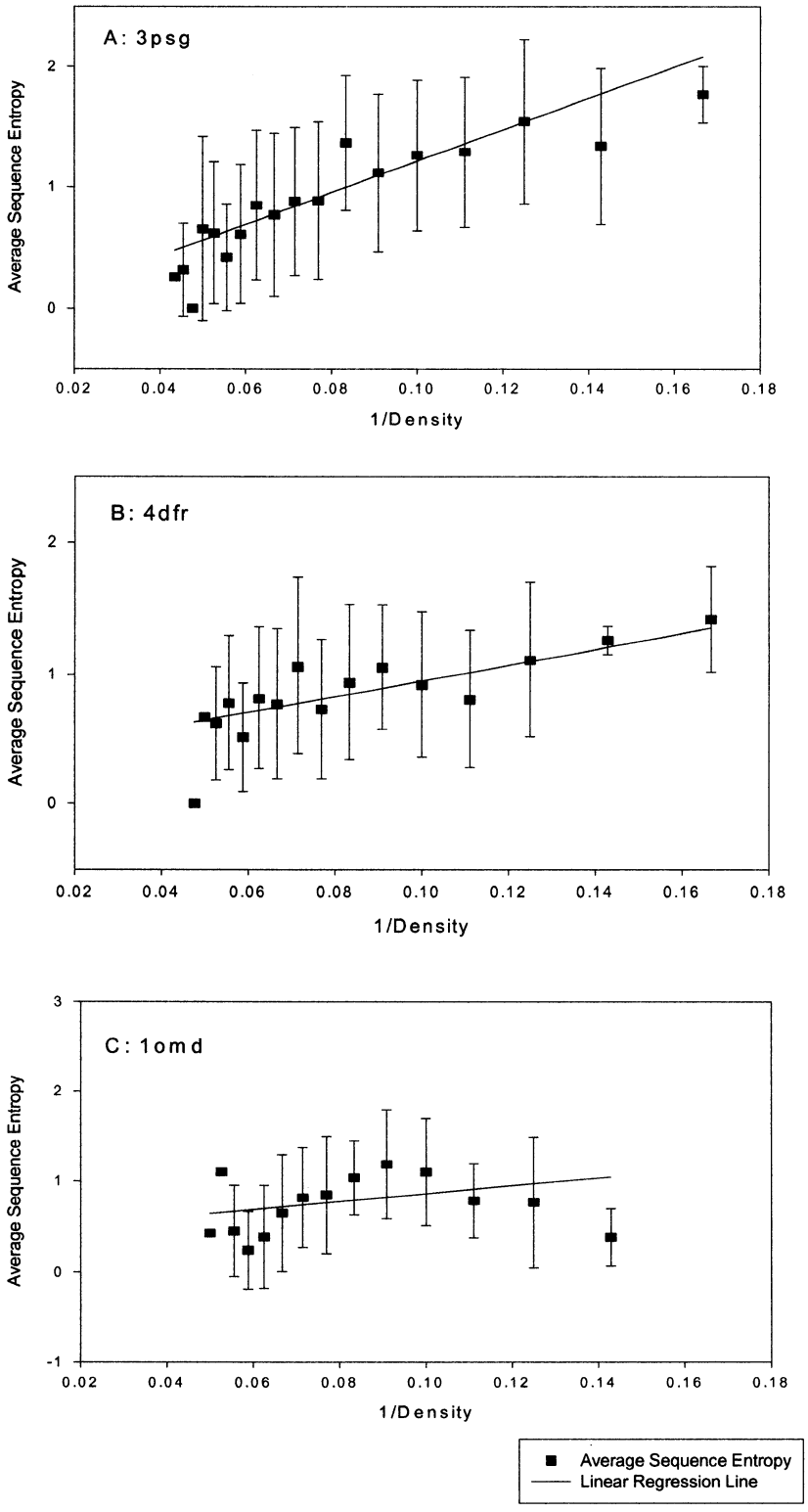
| PDB ID | r[A] | PDB ID | r | PDB ID | r | PDB ID | r |
|---|---|---|---|---|---|---|---|
| 1a1i[B] | 0.2456 | 1aq0 | 0.4368 | 1crc[BC] | 0.2298 | 3cna | 0.2766 |
| 1a1s[BC] | 0.1565 | 1aqh[B] | 0.4083 | 1crm | 0.3583 | 3est | 0.5359 |
| 1a32 | 0.248 | 1atp[B] | 0.3378 | 1crz | 0.2506 | 3gbp | 0.3856 |
| 1a3c[C] | 0.1661 | 1av5 | 0.3421 | 1csr | 0.2466 | 3grs | 0.2402 |
| 1a3s | 0.2506 | 1av6 | 0.3283 | 1d6m | 0.3509 | 3pfk | 0.4256 |
| 1a48 | 0.1944 | 1av7[B] | 0.4327 | 1daj | 0.3585 | 3pgk[B] | 0.3051 |
| 1a59[C] | 0.1367 | 1aw5[C] | 0.1487 | 1dcs[C] | 0.1382 | 3pgm | 0.3379 |
| 1a5z[C] | 0.1594 | 1aw9[C] | 0.0995 | 1dhs | 0.301 | 3psg[B] | 0.4734 |
| 1a6f[C] | 0.2057 | 1aye | 0.3134 | 1dht | 0.2992 | 3rn3[BC] | 0.1892 |
| 1a6q | 0.3555 | 1ayl | 0.377 | 1din | 0.3064 | 3rp2 | 0.5143 |
| 1aat | 0.2778 | 1ayx | 0.2456 | 1dmr | 0.3295 | 4ape | 0.3771 |
| 1ab4[B] | 0.278 | 1azi[BC] | 0.1378 | 1e1k | 0.3321 | 4dfr | 0.2729 |
| 1acb | 0.5297 | 1ba3[C] | 0.1179 | 1e3h | 0.1817 | 4mdh[B] | 0.2042 |
| 1add | 0.2177 | 1bc2[C] | 0.1924 | 1e3q | 0.3899 | 4pep[B] | 0.4316 |
| 1adi[B] | 0.3017 | 1bf2 | 0.2177 | 1e5m[B] | 0.3092 | 4tnc[BC] | 0.1245 |
| 1ae4 | 0.2786 | 1bfd | 0.2265 | 1ebv | 0.3143 | 5acn | 0.278 |
| 1af3[C] | 0.2358 | 1bg0 | 0.4307 | 1eeh | 0.2858 | 5cha | 0.5547 |
| 1agm | 0.2619 | 1bg3 | 0.311 | 1hgu[BC] | 0.1975 | 5cpa | 0.3942 |
| 1agx | 0.4187 | 1bia[C] | 0.1852 | 1lz1[C] | 0.3912 | 5cpv[C] | 0.1884 |
| 1aha[C] | 0.2579 | 1bit[B] | 0.3644 | 1omd[C] | 0.1432 | 5cts | 0.3208 |
| 1ahn | 0.2855 | 1blz | 0.3587 | 1rbp[C] | 0.2012 | 5ldh[B] | 0.3791 |
| 1ai2[B] | 0.2214 | 1bn6[C] | 0.1652 | 1rhd[C] | 0.1296 | 5rub[C] | 0.1158 |
| 1ak2[C] | 0.1857 | 1bo6 | 0.3303 | 1ton | 0.5052 | 6ldh[B] | 0.3401 |
| 1ako | 0.3762 | 1boh | 0.2045 | 2act[B] | 0.4511 | 6xia | 0.3335 |
| 1al8 | 0.2098 | 1bsi[B] | 0.3416 | 2cts | 0.3262 | 7api[B] | 0.2757 |
| 1alc[C] | 0.2698 | 1bt3 | 0.4383 | 2lbp | 0.2032 | 7cat[B] | 0.3094 |
| 1aln | 0.2823 | 1bul[B] | 0.3419 | 2ldx[B] | 0.3262 | 8adh[B] | 0.2775 |
| 1amn | 0.3562 | 1bxq | 0.3524 | 2liv | 0.1931 | 8atc | 0.2888 |
| 1amp | 0.209 | 1byt[C] | 0.1442 | 2prk | 0.4493 | 8dfr | 0.3271 |
| 1an9 | 0.2924 | 1cb0 | 0.2973 | 2rn2 | 0.2881 | 9pap[B] | 0.4607 |
| 1ang[C] | 0.2579 | 1cex | 0.2142 | 2taa | 0.2423 | 9wga | 0.4612 |
| 1ao5 | 0.4424 | 1cjx | 0.3782 | 3blm | 0.3277 | | |
| 1aob[B] | 0.2702 | 1ck6 | 0.3388 | 3cla[C] | 0.1414 | | |

[A] Note r is linear-correlation coefficient of the correlation between sequence entropy and inverse of packing density of each protein.

[B] Indicates default maximum of 100 alignments.

[C] Proteins with P values in range of 0.001 to 0.15 (all others $P<0.001$).

**Figure 2**. Correlation plots of sequence entropy and inverse of packing density for range of sample proteins. The inverse of packing density (abscissa) is calculated from the query protein's atomic coordinates, determining the number of residue's $C^{\alpha}$ atoms within a 9 Å radius. Sequence entropy is calculated from the aligned set generated by BLASTP from the query sequence, and average entropy (closed squares) is determined by averaging sequence entropy for all sequence positions within an interval of packing density. Error bars for sequence entropy are obtained with respect to packing density. A) For pepsinogen (3psg: 365 residues), straight-line fit for raw data is y=13.020x-0.088; the correlation coefficient is 0.447; P<0.001. Straight-line fit for average sequence entropy (line) is y=12.070x-0.086. B) For dihydrofolate reductase (4dfr: 158 residues), straight-line fit for raw data is y=6.064x+0.342; the correlation coefficient is 0.274; P<0.001. Straight-line fit for average sequence entropy is y=7.350x+0.22. C) For oncomodulin (1omd: 107 residues), straight-line fit for raw data is y=4.328x+0.427; the correlation coefficient is 0.141; P<0.15. Straight-line fit for average sequence entropy is y=1.624x+0.59.


A: 3psg
Average Sequence Entropy
2
1
0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
1/Density
B: 4dfr
Average Sequence Entropy
2
1
0
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
1/Density
C: 1omd
Average Sequence Entropy
3
2
1
0
-1
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
1/Density
Average Sequence Entropy
Linear Regression Line

straight-line fits of average sequence entropies with respect to inverse of packing density are y=12.070x-0.086, y=7.350x+0.22, y=1.624x+0.59, and respective correlation coefficients are 0.898, 0.796 and 0.149.

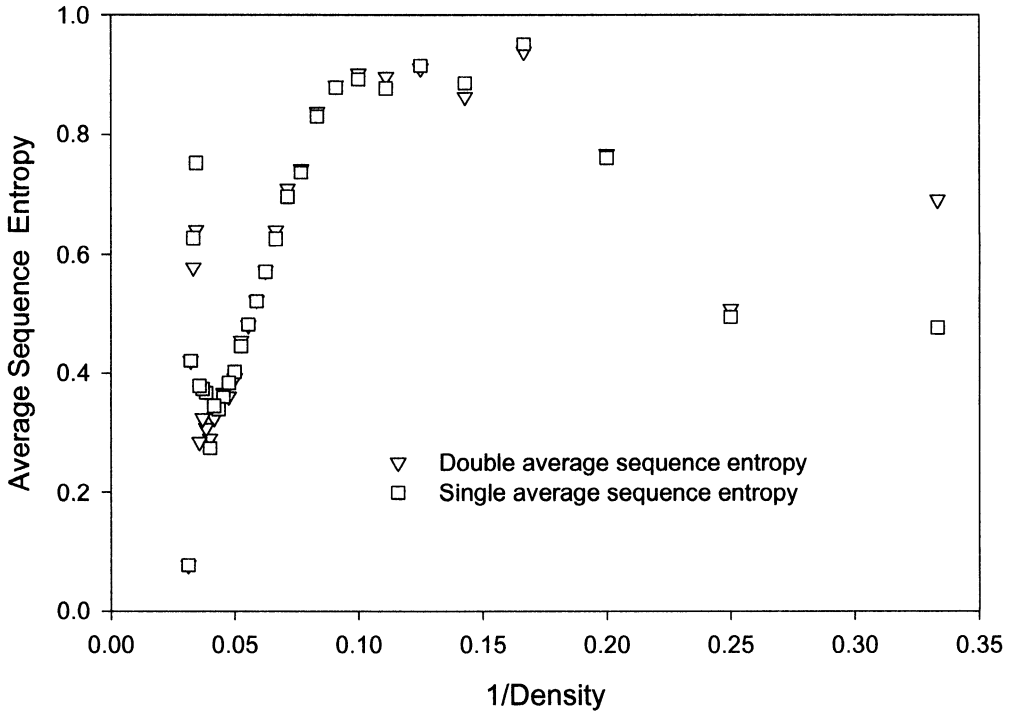
The aggregate (i.e., for all 130 alignment sets) average sequence entropies versus inverse $C^{\alpha}$ packing density correlation plots are shown in Figure 3. Here, single average entropy is done by summing individual residue entropies for a particular $C^{\alpha}$ packing density from all 130 protein sets of alignments. Double average entropy involves: First, calculating average entropy per density interval for individual proteins; Second, averaging again for all 130 proteins with respect to packing density. This is one way to investigate if there are appropriate ways to decrease noise at each interval of inverse density.

Four regions are observed in the Figure 3 plot:

I. A decrease in sequence entropy is noted starting from residues with greatest packing density of 35 and the least flexibility) to a density of 26 (0.03 and 0.038 of inverse of packing density).

II. The residues with inverse packing densities ranging from 0.04 to 0.083 (involving packing of 25 to 12 $C^{\alpha}$ atoms

**Figure 3**. Correlation plots of aggregate average sequence entropy of 130 protein alignment sets and inverse of packing density for corresponding query proteins. Here inverse packing density (abscissa) is calculated as noted in Figure 2 and aggregate sequence entropy (ordinate) is calculated in two ways: Single averaging (open squares) and double averaging (open triangles). Single averaged sequence entropy is determined by summing sequence entropy for each associated residue position with respect to the interval of inverse of packing density (abscissa). Double averaged sequence entropy is calculated by summing each protein's average entropy for a particular density interval.


1.0
0.8
0.6
0.4
0.2
0.0
Average Sequence Entropy
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
1/Density
Double average sequence entropy
Single average sequence entropy

within 9 Å radius), showing a proportional increase in sequence entropy.

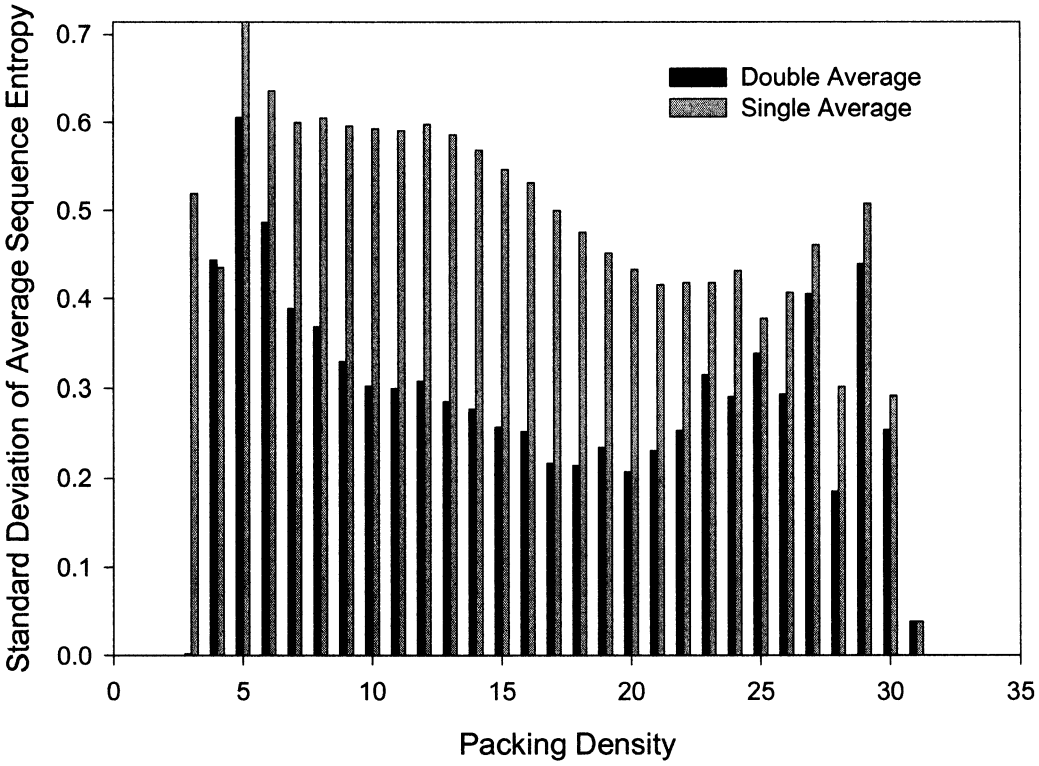
III. A flattening in sequence entropy values is found between 0.09 and 0.17 of inverse of packing density (involving 11 to 6 $C^{\alpha}$ atoms).

IV. Another decrease in sequence entropy is noted for the most flexible residues ranging from a packing density of 5 to 3 (0.20 and 0.33 of inverse of packing density).
In fact, the basic features for these regions can be found in most of 130 individual protein's correlation plots of average sequence entropy with respect to inverse packing density.

Figure 4 shows the standard deviations with respect to each packing density. It shows that double average entropy significantly decreases the standard deviation compared to single average sequence entropy. The average standard deviation for single average entropy is 0.508, and for double average sequence entropy is 0.341. However, except for the reductions in standard deviation with double averaging, the two types of averaged sequence entropy are effectively identical.

As shown in Figure 1A, there are a total of 41,632 residues for 130 proteins, and the region of packing of 25

**Figure 4**. Standard deviation of average entropy with respect to packing density. The estimated average standard deviation is 0.5 for single averaged entropy and 0.3 for double averaged entropy.

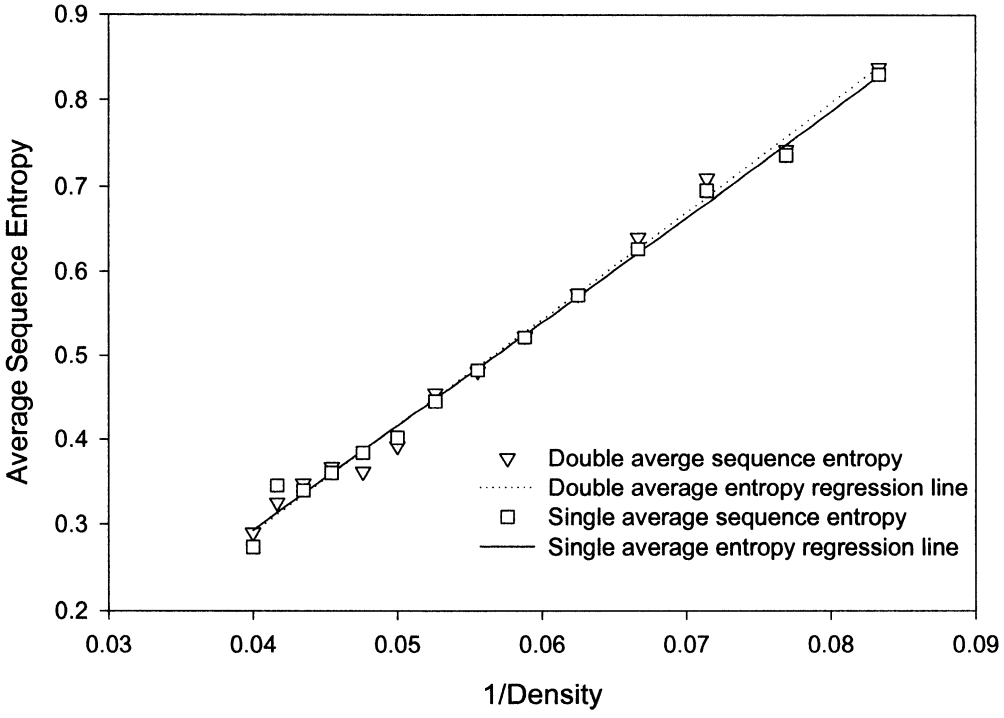

Standard Deviation of Average Sequence Entropy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
35
Packing Density
Double Average
Single Average

to 12 $C^{\alpha}$ atoms within 9 Å radius has 31,169 residues, that is 74.9 percent of total residues. The region between 6 to 11 $C^{\alpha}$ atoms has 10,173 residues, involving 24.4 percent of the aligned residues. The region between 3 to 5 $C^{\alpha}$ atoms has 138 residues (0.33 percent of total residues) and the region between 26 to 35 $C^{\alpha}$ atoms has 152 residues (0.36 percent of total residues).

The dominant region (packing of 25 to 12 $C^{\alpha}$ atoms within 9 Å radius), II, for the aggregate correlation plots is shown in Figure 5. Here the single average and double average sequence entropy are shown to be linearly correlated with respect to inverse packing density. The straight-line fit for the aggregate single average sequence entropy versus inverse packing density is y=12.350x-0.20; correlation coefficient is 0.997; P<0.001. The straight-line fit involving aggregate double entropy is effectively identical: y=12.658x-0.22; correlation coefficient is 0.997; P<0.001.

Shown in Figure 6 are overlays of three sets of averaged hydrophobicities and single averaged sequence entropy as a function of inverse packing density. Residue hydrophobicity is obtained for every query protein residue that is part of an alignment using three different scales:

**Figure 5**. Linear regression of selected regions for correlation plots involving 31,169 aligned query residues out of a total of 41,632. Included from Figure 3 are aggregate averaged sequence entropy values (ordinate) corresponding to region of inverse packing density (abscissa) of 0.040 to 0.083 (corresponding to packing density of 25 to 12 $C^{\alpha}$ atoms within a 9 Å radius). Aggregate single averaged entropy (open squares) straight-line fit (line) is y=12.350x-0.20; correlation coefficient is 0.997; P<0.001. Aggregate double averaged entropy (open triangle) straight-line fit (dotted line) is y=12.658x-0.22; correlation coefficient is 0.997; P<0.001.

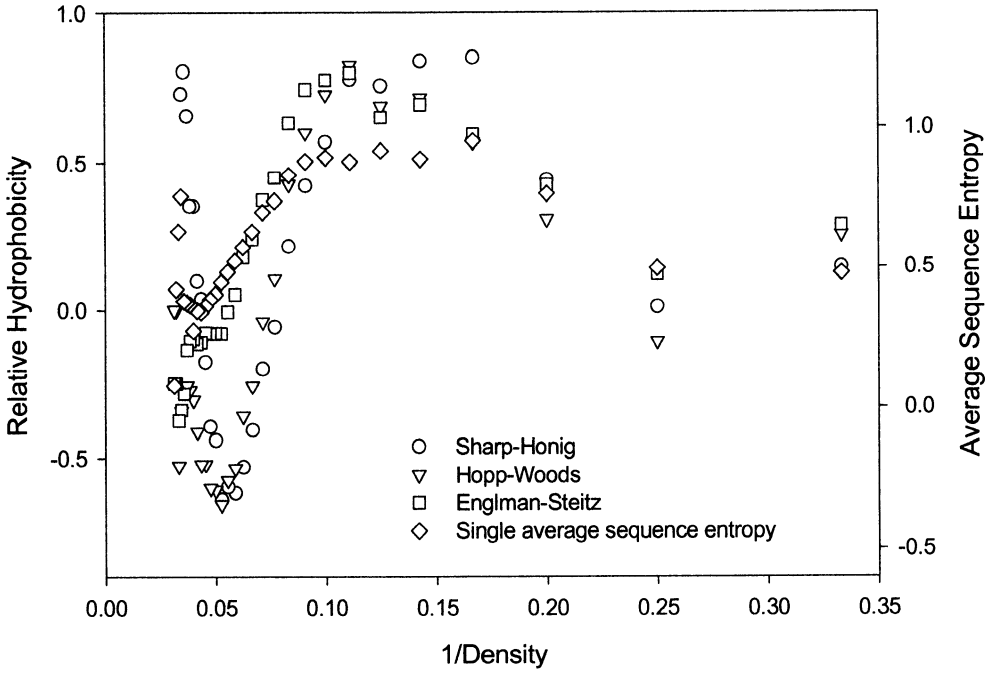

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Average Sequence Entropy
0.03
0.04
0.05
0.06
0.07
0.08
0.09
1/Density
Double averge sequence entropy
Double average entropy regression line
Single average sequence entropy
Single average entropy regression line

Sharp-Honig, Hopp-Woods, and Engleman-Steitz (Sharp et al., 1991; Hopp & Woods, 1981; Engleman & Steitz, 1986). For each scale, an averaged hydrophobicity is calculated from the set of all residues within a particular density interval. Averaged Engleman-Steitz scales are normalized by multiplying a constant of 0.25 to their average hydrophobicity values. Sharp-Honig scales are normalized by the formula (-x+3).

Clearly, all three sets of query hydrophobicity values (Figure 6) share regions of similarity with the corresponding sequence entropy values: a dominant region (region II) in which hydrophobicity increases as a function of inverse of packing density, a flattening region (region III), and two flanking regions (region I & IV) involving residues with low and high flexibility in which relative hydrophobicity energies decrease. The dominant region (region II) shows significant linear correlations for all three hydrophobicity sets. Though there are shifts in the breakpoints with hydrophobicity plots for the four regions.

Region III (flattening of sequence entropy values) suggests there is a large fraction of non-hydrophobic residues embedded in regions that are probably accessible to water. The minimum value for the three normalized

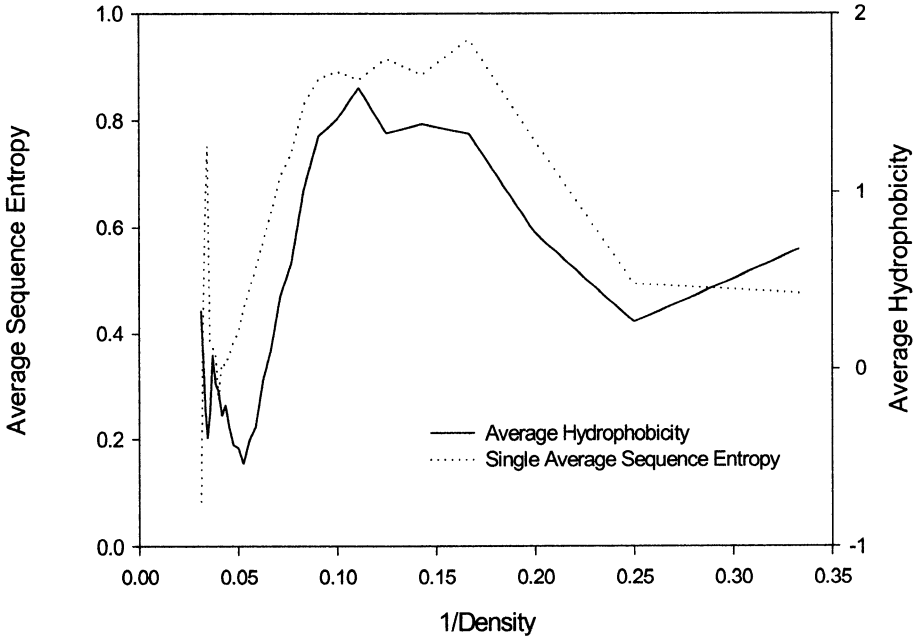
**Figure 6**. Overlay of three sets of average hydrophobicity per residue and aggregate single-averaged sequence entropy with respect to inverse $C^{\alpha}$ packing density. Residue hydrophobicity (left-ordinate) is calculated for each query protein, weighting the aligned residue type with different scales: Sharp, Honig and coworkers, 1991 (circles); Hopp and Woods, 1981 (inverted triangles); Engleman, Steitz and coworkers, 1986 (square). The average hydrophobicity for each scale is calculated by summing the residue hydrophobicities for all aligned residues found within an interval of packing density. Engleman-Steitz scales are normalized by multiplying a constant of 0.25 to their average hydrophobicity values; Sharp-Honig scales are normalized by the formula (-x+3). Aggregate single-averaged sequence entropy (diamonds) is labelled on right-ordinate axis.


1.0
0.5
0.0
-0.5
Relative Hydrophobicity
1.0
0.5
0.0
-0.5
Average Sequence Entropy
Sharp-Honig
Hopp-Woods
Englman-Steitz
Single average sequence entropy
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
1/Density

hydrophobicities (Figure 6) calculated from representative hydrophobicity scales, vary over an inverse density range of 0.040 to 0.050. Though less variation is noted for the breakpoint between region I and II, between inverse densities 0.083 and 0.091, such differences suggest some uncertainties in certain experimental measurements and calculations for hydrophobicity scales. The minimum value and breakpoint for the curves of the single averaged sequence entropy and the three normalized hydrophobicity plots calculated from different scales remain consistent.

Shown in Figure 7 is an overlay of average of three sets of normalized hydrophobicities and single averaged sequence entropy as a function of inverse packing density. The two lines show highly similarity even in high and low density regions. As noted in Figure 6 the original three sets of hydrophobicity values are among themselves very similar.

**Figure 7.** Overlay of average of three sets of average hydrophobicity per residue and aggregate single-averaged sequence entropy with respect to inverse $C^{\alpha}$ packing density. Average hydrophobicity (left-ordinate) is calculated by averaging Sharp-Honig, Hopp-Woods, and Engleman-Steitz hydrophobicity values shown in Figure 6.


Average Sequence Entropy
Average Hydrophobicity
1/Density
Average Hydrophobicity
Single Average Sequence Entropy
1.0
0.8
0.6
0.4
0.2
0.0
2
1
0
-1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35

# DISCUSSION

## Flexibility and sequence entropy

Here a correlation relationship between information sequence entropy and $C^{\alpha}$ packing for 130 proteins, ranging in size from 85 to 901 residues, was determined. It has been found that protein local flexibility is strongly correlated with inverse of $C^{\alpha}$ packing (Bahar et al., 1997). This flexibility can be experimentally correlated to an entropic term involving the change of the free energy for a protein under deformation (Bahar et al., 1998). Regions of low sequence variability can appear resistant to unfolding. Here residues' inverse $C^{\alpha}$ packing density is calculated from X-ray structure in the PDB databank for each of the 130 query proteins. Then sequence entropy, in the form of the Shannon entropy partitioned by residue, is calculated from BLASTP sequence alignment.

The major regions are identified when plotting sequence entropy with respect to inverse packing density (Figure 3 & Figure 5), where more than 74.9 percent of sequence positions exhibit a strong linear dependence inclusive of a range of 0.040 through 0.083 inverse $C^{\alpha}$ packing density per 9 Å radius (region II). There is a strong correlation

between sequence variability and local amino acid flexibility. This correlation quantitates the previous qualitative relationship between the degree of residue burial and sequence variability. Sequence entropy is a Shannon entropy that is a scaled expression for sequence variability. The correlation also confirms that Shannon-based sequence entropy is consistent with other expressions for conformational entropy as was shown by Saven and coworkers (Zou & Saven, 2000; Kono & Saven, 2001).

The correlation between sequence variability and local amino acid flexibility is consistent with a similar pattern noted by Lustig and coworkers with respect to peptide binding to RNA (Hsieh et al., 2002). Enhanced flexibility at a particular residue position is associated with the ability of local structure to accommodate mutation. Such a description can more broadly be related to sequence positions in a folded protein, where the ability to accommodate mutation corresponds to allowing a range of flexible three dimensional features, such as loops, and possibly alternative residue contacts. However key structural features associated with conserved regions of structure are maintained.

In addition, another 24.4 percent of sequence positions (region III) indicate some relative minimum of both sequence entropy and the percent strongly hydrophobic residue type at inverse densities above 0.083. This suggests some 10 percent of strongly hydrophobic residues (Liao et al.), I, F, L, M, V, W, and Y (Poupon & Mornon, 1999), allowed in region III that may be accessible to water.

The linear correlation observed for region II involves average behavior because the average sequence entropy is summed over all residues with respect to inverse of packing density. For the single residues involving individual proteins, the plots show extensive noise. This suggests additional methods need to be developed in regards to the calculation of individual entropies.

Anomalous behavior is found above packing densities of 26 $C^{\alpha}$ per 9 Å radius and below densities of 5 (region I & IV). Since the region between 26 to 35 $C^{\alpha}$ atoms and the region between 5 and 3 include only 152 and 138 query residues, respectively, out of a total of 41,632, they are not universal. In addition less than 60 percent of the 130 query proteins are sampled in either anomalous region.

Interestingly, region I shows a disproportionate weighting in alanine (A) and glycine (G) residues (Lustig &

Yeh, unpublished work). This is consistent with the notion that in region I there is a relative bias in the number of these small residues, whose additional flexibility is not accounted for by calculations of inverse $C^{\alpha}$ density. Region IV shows a disproportionate fraction of gap residues. One must be careful in assessing the importance of the two flanking regions I and IV. Their populations are very small and unrepresentative. In addition they may be, at least in part, an artifact of error with respect to the original X-ray coordinates.

**Sequence entropy and local flexibility calculation**

Koehl and Levitt introduced an approach to quantify the sequence entropy with respect to packing (Koehl & Levitt, 2002). The sequence information contained in a multiple sequence alignment is converted into a profile matrix, which consists of an array of vectors, one for each position in the sequence. In their method, the entropy calculation is complex and only 10 proteins have been used to investigate the relationship between sequence entropy and protein structure. In this research, one can simplify the calculation and use a large set of 130 query proteins.

In addition, Koehl and Levitt calculated an entropy for residues from structural physical parameters such as geometry of protein backbone. Interestingly, for a small set of proteins, a linear correlation was observed between this entropy and sequence entropy.

**Hydrophobicity and sequence variability**

Hydrophobicity is one of the most important physical-chemical/biological properties of individual amino acids with respect to their propensity to be miscible in water. The calculated average sequence entropy and hydrophobicity values from various measures of aggregate hydrophobicity for all 130 proteins show a strong similarity (Figure 7), even in the anomalous regions of high and low density, with respect to inverse $C^{\alpha}$ packing density. All of this suggests that for most residue positions, as noted in region II, the ability to accommodate mutation as measured by sequence entropy is inversely correlated to degree of packing. On average, for a particular amino acid type, hydrophobicity can be correlated to this degree of residue packing.

Hydrophobicity is a measure of the degree of amino acid burial within a protein, so it is not a surprise that hydrophobicity values correspond to sequence entropy. For

each protein, sequence entropy is calculated at each query sequence residue position. It is a function of the sequence variability as determined by BLASTP at that position. The hydrophobicity for each residue position of the query sequence is calculated by a simple metric weighting of each amino acid type.

Again, the sequence entropy and the hydrophobicity calculations include averaging with respect to all residues corresponding to some interval of density. So clearly the correlations between the log scaled sequence variabilities, indicating sequence entropy, and the corresponding hydrophobicity are consistent with some averaged behavior for residues of a given packing density.

Still, this sort of correlation between sequence entropy and hydrophobicity is interesting given the latter's critical importance in the correct folding of model protein chains (Hinds & Levitt, 1994; Dill, 1995). This is consistent with core hydrophobic residues often being described as conserved and where generally the degree of amino acid burial correlates well with hydrophobicity scales (Miyazawa & Jernigan, 1993).

The mutability, as measured by sequence entropy, is inversely correlated with a degree of residue packing. The

propensity for packing of a particular amino acid type is, in aggregate, representative of its hydrophobicity. Moreover, hydrophobicity scales are correlated with amino acid charge and volume characteristics for the various classes of amino acids (Pickett & Sternberg, 1993).

Notably, average energies for the various amino acid types calculated from the frequency of residue contacts in intra-protein and protein-protein interactions correlate well with existing hydrophobicity scales (Miyazawa & Jernigan, 1993; Young et al., 1994). In regards to the reverse situation, hydrophobicity based scoring matrices show strong fidelity in sequence alignment scoring when compared with the more conventional PAM and BLOSUM matrices (Vogt et al., 1995). However, as we have shown with the various classes of sequence entropy behavior, sequence variability by itself is not a unique descriptor for hydrophobicity, given that there is not always a unique one to one relationship between relative hydrophobicity and inverse density. However the three sets of hydrophobicities show minor variability in breakpoints with respect to density, the average of these values is highly correlated to the corresponding average sequence entropy.

## CONCLUSION

Two major regions are identified when plotting sequence entropy with respect to inverse packing density, where more than 74 percent of sequence positions exhibit a linear dependence inclusive of a range of 0.040 through 0.083 inverse $C^{\alpha}$ packing density per 9 Å radius. In addition, another 24 percent of sequence positions indicate some relative minimum sequence entropy and percent strongly hydrophobic residue type at inverse densities above 0.083. This suggests some minimum number of strongly hydrophobic residues allowed in regions that may be accessible to water.

Interestingly, various measures of aggregate hydrophobicity for all 130 proteins overlay reasonably well with all four regions identified by correlation plots of sequence entropy versus inverse packing density. All of this suggests that for most residue positions, the ability to accommodate mutation as measured by sequence entropy is inversely correlated to their degree of packing. On average for a particular amino acid type, hydrophobicity can be correlated to this degree of residue packing. Further understanding the structural connections with sequence entropy remains of great interest.

## FUTURE STUDIES

Further understanding the protein structural connections with sequence entropy remains of great interest and should include:

1. Exploring the structural composition of query residues in the two anomalous regions (high density region and low density region). Determining their secondary structure and the substitution pattern of corresponding aligned residues.
2. Characterizing the role of water, including surface and interior cavities.
3. Exploring the nature of packing density (e.g. all-atom or calculated excluded volume), including in the high density and low density regions.

## REFERENCES


Altschul, S. F.; Boguski, M. S.; Gish, W.; Wooten, J. C. Issues in searching molecular sequence databases. Nature Genetics 1994, 6, 119-129.

Altschul, S. F.; Gish, W.; Miller, W., Myers, E.; Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. 1990, 215, 403-410.

Altschul, S. F.; Gish, W. Local alignment statistics. Meth. Enzymol. 1996, 26, 460-480.

Altschul, S. F.; Madden, T. L.; Scaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 1997, 25, 3389-3402.

Bahar, I.; Ali, R. A.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Folding and Design 1997, 2, 173-181.

Bahar, I.; Wallqvist, A.; Covell, D. G.; Jernigan, R. L. Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model. Biochemistry 1998, 37, 1067-1075.

Baker, D.; Salij, A. Protein structure prediction and structural genomics. Science 2001, 294, 93-96.

Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 1977, 112, 535-542.

Bevington, P. R. Data Reduction and Error Analysis for the Physical Sciences; McGraw-Hill: New York, 1969.

Brown, T. A. Genomes; Wiley: New York, 2002.

Creighton, T. Proteins: Structures and Molecular Properties; Freeman Press: New York, 1994.

Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. Atlas of protein sequence and structure. National Biomedical Research Foundation 1978, 5, Suppl. 3. 345-352.

Dill, K.; Bromberg, S; Yue, K.; Feibig, K.; Yee, D.; Thomas, P.; Chan, H. Principles of protein folding - a perspective from simple exact models. Protein Sci. 1995, 4, 561-602.

Engleman, D. M.; Steitz, T. A. The spontaneous insertion of proteins into and across membranes: the helical hairpin hypothesis. Cell 1981, 23, 411-422.

Engleman, D. M.; Steitz, T. A.; Goldmann, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Chem. 1986, 115, 321-353.

Fauchere, J. L.; Pliska, V. Hydrophobic Parameters $\pi$ of amino acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. 1983, 18, 369-375.

Fischer, D.; Eisenberg, D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc. Natl. Acad. Sci. USA 1997, 94, 11929-11934.

Gross, E. A.; Li, G. R.; Lin, Z. Y.; Ruuska, S. E.; Boatright, J. H.; Mian, I. S.; Nickerson, J. M. Prediction of structural and functional relationships of Repeat 1 of human interphotoreceptor retinoid-binding protein (IRBP) with other proteins. Mol. Vis. 2000, 6, 30-39.

Hinds, D. A.; Levitt, M. Exploring conformational space with a simple lattice model for protein structure. J. Mol. Biol. 1994, 243, 668-682.

Hopp T. P.; Woods K. R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Nat. Acad. Sci. USA 1981, 78, 3824-3828.

Hsieh, M.; Collins, E. D.; Blomquist, T.; Lustig, B. Flexibility of BIV TAR-Tat: Models of peptide binding. J. Biomol. Struct. Dyn. 2002, 20, 243-251.

Jones, D. T. Protein structure prediction in the postgenomic era. Curr. Opinion Struct. Biol. 2000, 10, 371-379.

Karlin, S.; Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 1990, 87, 2264-8.

Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 1959, 14, 1-63.

Koehl, P.; Levitt, M. Protein topology and stability define the space of allowed sequences. Proc. Natl. Acad. Sci. USA 2002, 99, 1280-1285.

Kono, H.; Saven, J. G. Statistical theory for protein combinatorial libraries. J. Mol. Biol. 2001, 306, 607-628.

Larson, S. M.; England, J. L.; Desjarlais, J. R.; Pande, V. S. Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Sci. 2002, 11, 2804-2813.

Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 1976, 104, 59-107.

Liao, H.; Yhe, W.; Chiang, O.; Jernigan, R. L.; Lustig, B. Protein sequence entropy is closely related to paking density and hydrophobicity. Prot. Engr., submitted for publication, 2004.

Lustig, B.; Bahar, I; Jernigan, R. L. RNA bulge entropies in the unbound state correlate with peptide binding strengths for HIV-1 and BIV TAR RNA because of improved conformational access. Nucl. Acids Res. 1998, 26, 5212-5217.

Lustig, B.; Yeh, W. San Jose State University, San Jose, CA. Unpublished work, 2004.

Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. Ann. Rev. Biophys. Biomolec. Struct. 2000, 29, 291-325.

Miyazawa, S.; Jernigan, R. L. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Prot. Engr. 1993, 6, 267-278.

National Center for Biotechnology Information (2002): http://www.ncbi.nlm.nih.gov/

Papp, P. P.; Chattraj, D. K.; Schneider, T. D. Information analysis of sequences that bind the replication Initiator RepA. J. Mol. Biol. 1993, 233, 219-230.

Pickett, S. D.; Sternberg, M. J. E. Empirical scale of side-chain conformational entropy in protein folding. J. Mol. Biol. 1993, 231, 825-839.

Pilpel Y.; Lancet, D. The variable and conserved interfaces of modeled olfactory receptor proteins. Protein Sci. 1999, 8, 969-977.

Poupon, A.; Mornon, J. P. "Topohydrophobic positions" as key markers of globular protein folds. Theor. Chem. Acc. 1999, 101, 2-8.

Protein Data Bank (2002):http//www.rcsb.org.pdb/

Scalley-Kim, M.; Minard, P.; Baker D. Low free energy cost of very long loop insertions in proteins. Protein Sci. 2003, 12, 197-206.

Schneider, T. D.; Stormo, G. D.; Gold, L. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 1986, 188, 415-431.

Sharp, K. A.; Nicholls, A.; Friedman, R.; Honig B. Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. Biochemistry 1991, 30, 9686-9697.

Valdar, W. S. J. Scoring residue conservation. Proteins 2002, 48, 227-241.

Vogt, G.; Etzold, T.; Agros, P. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. J. Mol. Biol. 1995, 249, 816-831.

Young, L.; Jernigan, R. L.; Covell, D. G. A role for surface hydrophobicity in protein-protein recognition. Protein Sci. 1994, 3, 717-729.

Zou, J.; Saven, J. G. Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. J. Mol. Biol. 2000, 296, 281-294.